Built to evolve

B. Hoeneisen and G. Trueba

Universidad San Francisco de Quito
15 June 2005

Abstract

We study the probabilities of evolution based on random mutations and natural selection. We conclude that evolution to multicellular eukaryots, or even prokaryots, is unlikely to be the result of only random mutations. Complex organisms have evolved through several mechanisms besides random mutations, namely DNA recombination, adaptive mutations, and acquisition of foreign DNA. We conclude that all living organisms, in addition to being self-organizing and reproducing (autopoyetic), have built-in mechanisms of evolution, some of which respond in very specific ways to environmental stress.

1 Introduction

We consider the probabilities of obtaining genomes by random mutations and natural selection, and obtain the conditions that are required for successfull evolution. Examples are presented for viruses, bacteria, eukaryote cells, and multi-cellular organisms. Our conclusions are collected in Section 10.

2 The odds

The human genome has about 30 thousand useful genes, each with an average coding region of 20 thousand base pairs coding about 6666 aminoacids.[1] What is the probability of writing a specific sequence of 30000×6666 words (aminoacids) chosen at random from a list of 20? The answer is $20^{-30000\times 6666} \approx 10^{-260000000}$, i.e. zero for all practical purposes. For that matter, what is the probability of getting the correct genome of any one of the 30 million species, with one try every second since the Big Bang, by every one of the 10^{31} (or so) bacteria on Earth? The answer is $3 \times 10^7 \times 4 \cdot 10^{17} \times 10^{31} \times 20^{-30000\times 6666}$

which is still the same result: $\approx 10^{-260000000}$ (only the last two digits in the exponent are changed), i.e. zero for all practical purposes.

Note that the "simplest" living organism, the prokaryote bacteria, is almost as complex as a human being: it has of order 1000 active genes with an average of ≈ 1500 base pairs! There are many missing links between organic molecules (sugars, lipids, bases, aminoacids, etc) and the simplest forms of life.

Perhaps the trick is to write little pieces of genetic code at a time. So, let us turn the question around: What is the largest gene that can be produced at random with a finite probability (say, 10^{-4}) with one try every second during 100 million years by each of the 10^{31} bacteria on Earth? The answer is one gene encoding, at most, ≈ 39 aminoacids, or ≈ 117 base pairs. This is roughly the limit for undirected random evolution.

3 Evolution

Let us play a game called "Evolution". We sit in front of a key board and hit keys at random. The probability of obtaining "Romeo and Juliet" is zero for all practical purposes. In fact, the probability of obtaining any meaningfull novel in any known language is zero for all practical purposes. Now introduce "mutations", i.e. replace random letters by random hits of the keyboard. Still no meaningfull novel will ever be obtained for all practical purposes.

Now suppose we are allowed to select which letter (or small set of letters) to mutate at a rate much higher than the background mutation rate of the other letters, and we are allowed to stop the hypermutations when a particular outcome is obtained. For example, choose to hypermutate the 10th letter until "e" is obtained, or choose to hypermutate the 10th, 11th and 12th letters until "dog" is obtained. Of course, now we can write any novel at all: the game has become trivial. If the choice of which letter to hypermutate is perfectly specific, and the choice of which outcome to select is perfectly specific, the game becomes trivial, i.e. the outcome becomes certain, even tho the keys are hit at random. If the environment were perfectly specific, it would have perfect control over evolution (even if mutations occur at random).

Evolution lies somewhere in between. The choice of which bases on which genes to hypermutate is not perfectly specific and is incomplete, and the outcome that stops the hypermutations is also not perfectly specific. However, with enough specificity it is possible to write little pieces of survivable genome. Then the pieces can be combined. In our example we could copy "dog", reverse "dog", concatenate "dog" and "cat", interchange "dog" and

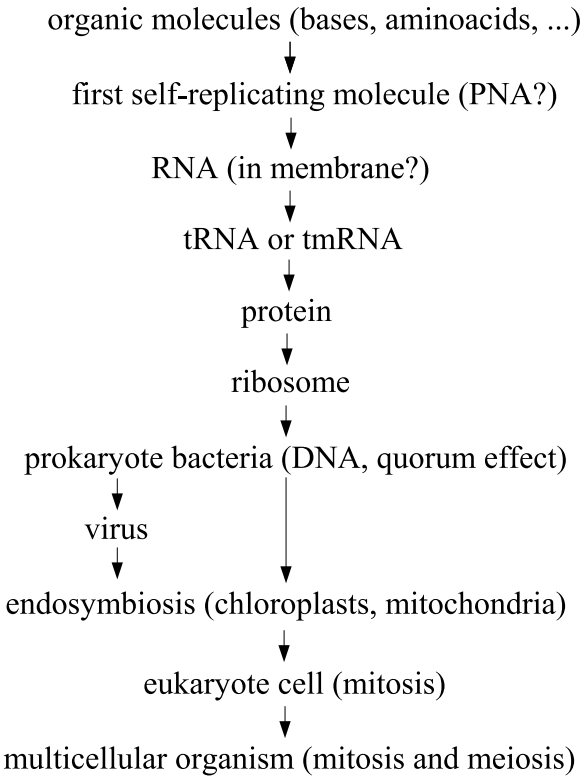

Figure 1: Plausible steps of evolution. For PNA see [2], for tRNA see [3, 4], for genetic recombination see [5], for viruses see [6].

"cat" (as in horizontal gene transfer, or in meiosis), introduce words and phrases from other books, and so on. The interchange of genetic material within, and between, bacteria, viruses and eukaryote cells plays a major role in evolutionary change.[6]

Survivable pieces of the genome will often involve negative (stabilizing) feedback loops. A hierarchy of negative feedback loops, within negative feedback loops, within ..., is self organizing.

The steps in evolution, from simple to complex, might have been as shown in Figure 1. At all levels of complexity the environment must have directed evolution with sufficient specificity for the steps to have a non-zero probability.

4 Adaptive mutations

Consider the (simplified) metabolic pathway of a cell shown in Figure 2. Precursor A is converted to end-product B by enzyme C. Enzyme C is encoded

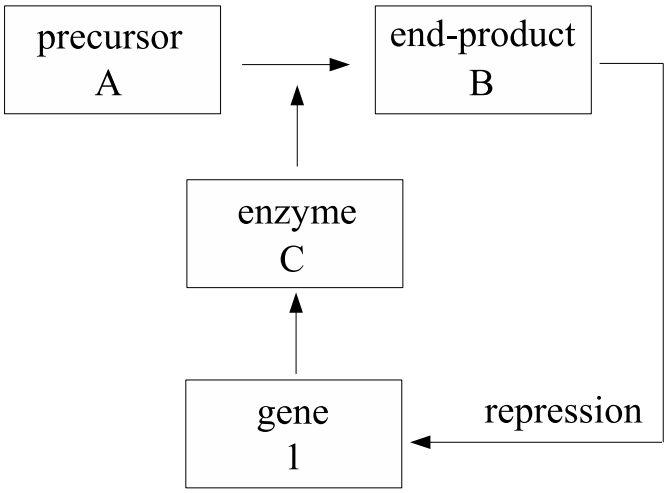

Figure 2: Precursor A is converted to end-product B by enzyme C. Enzyme C is encoded by gene 1. The concentration of B is regulated by the repression of gene 1 by the end-product B. Note that enzyme C is synthesized only when needed. This circuit is regulating and is built to evolve.

by gene 1. The transcription of gene 1 (and synthesis of enzyme C) is repressed (directly or indirectly) by the end-product B. This negative feedback loop regulates the concentration of B, and is efficient because enzyme C is produced only when needed. This feedback loop therefore has a value for survival, and would have been selected by nature. Now starve the cell of precursor A in the presence of a similar precursor A′. The result is a reduction of the concentration of the end-product B, and a derepression of gene 1. The rate of transcription of gene 1 increases (by a factor that can exceed 1000[7]). During transcription, mRNA copies one strand of DNA exposing the other strand. Single strand DNA is prone to mutations due to the lack of hydrogen bond stabilization between complementary bases, the formation of loops and other secondary structures, [1] and supercoiling.[7] As a result, gene 1 acquires a high rate of mutations and begins synthesizing enzymes similar to C, until one of them is able to convert precursor A′ into end-product B.

[1]Segments of the single strand DNA may stick to other segments with mostly complementary bases, resulting in unpaired or mispaired bases. Unpaired bases are prone to deamination, deletion or replacement. Cytosine deaminates to uracil at a rate 100 times larger in single strand DNA than in double strand DNA.[7] Mutations also occur in the end-loops where bases have no complement, and in the stem. These errors are immortalized during DNA duplication or repair.

Gene 1 is then repressed by end-product B and hypermutation stops. The resulting negative feedback loop resumes control of the concentration of the end-product B to the same original concentration. The net result of these processes is that a change of the environment (the starvation for precursor A) triggers mutations of a specific gene of the cell, until the cell is capable of substituting precursor A by precursor A′. So the environment can direct evolution in very specific ways.

Let us briefly describe examples.

5 B-lymphocytes

Let us consider B-lymphocytes of the immune system (see Figure 3), which have been studied in considerable detail.[8, 9] These B-lymphocytes have antibody proteins (called immunoglobulins) attached to their membrane. An invading bacteria has antigen proteins attached to its membrane. The antibody can bind to a very specific set of antigens. This binding triggers a series of complex steps (including helper T-lymphocytes) that activate mitosis of the B-cell, expresses several genes that code immunoglobulins, and differentiates the B-lymphocytes into antibody secreting cells and memory cells. Proliferanting B-cells in germinal centers show high rates of point mutations in genes coding immunoglobulins (10^3 to 10^6 times higher than the spontaneous rate of other genes).[9] The result is the synthesis of immunoglobulins with small differences. The binding of these antibodies to the antigens (with the intervention of follicular dendritic cells) produce signals that rescue the B-lymphocytes from programmed cell death. At the latter stages of the infection the concentration of invading bacteria becomes low, so only B-cells producing very high affinity antibodies can bind to the antigens and survive. This phenomenon is called "affinity maturation".

The net result of these complex processes is that a change of the environment (the invading bacteria) triggers hypermutations of specific sections (those that code for the binding site of the immunoglobulins) of specific genes of specific B-cells, and selects mutations producing immunoglobulins with the highest affinity to the antigen. Note that each one of these steps is very specific.

Since B-lymphocytes are somatic, the selected mutations are not passed on to the next generation, so this is an example of evolution of B-cells in one individual, not evolution of the species.

The hypermutation associated to affinity maturation of B-cells is caused by the induction of mutagenic genes such as cytidine deaminase (which causes C to U transitions) and error prone DNA polymerases. The presence of

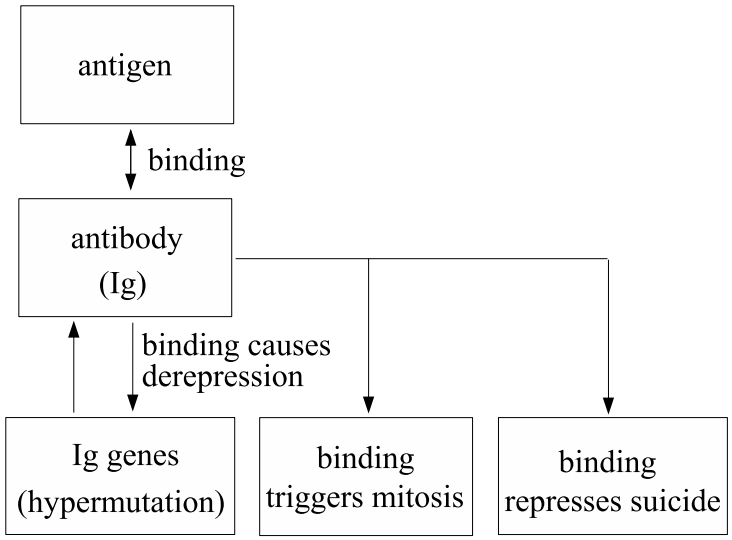

Figure 3: B-lymphocytes have antibody proteins (called immunoglobulins, Ig) attached to their membrane. Binding of these proteins to specific antigens triggers mitosis of the B-cell. Hypermutation of the Ig coding genes causes variation in the antibodies. Binding to the antigens rescues the B-cell from programmed death. The result is survival of the B-cells producing the most specific antibody.

the same genes in other eukaryots may indicate that similar hypermutagenic processes may occur in eukarya, and, as in prokaryots or B-cells, the induction of these genes may be triggered by environmental stress such as starvation.

Let us mention that B-lymphocytes with hypermutation and recombination of V, D and J cassettes of genes coding immunoglobulins, produce B-cells that synthesize of order 10^{11} different antibodies capable of binding to as many different antigens. So, with just a handfull of genes in the genome, B-cells are able to synthesize a much larger number of different proteins!

6 *Enterobacter arogenes*

Let us briefly describe the metabolic pathway of *Enterobacter arogenes* shown in Figure 4.[10, 11, 7] In the wild strain 5P14, ribitol induces gene 1 to synthesize ribitol dehydrogenase. This enzyme metabolizes ribitol, or, with low specific activity, xylitol. The wild bacteria can therefore metabolize xylitol only if ribitol is present. If the bacteria is starved for ribitol in the presence of xylitol, a mutation occurs in a gene 2 that causes the expression of gene 1 even in the absence of ribitol. This strain, called X1, appears in

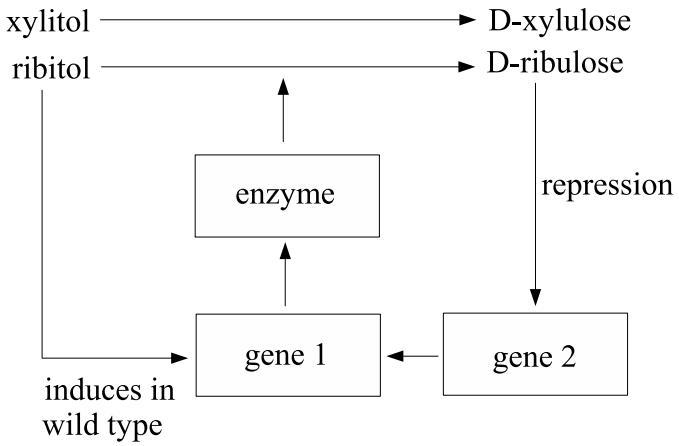

Figure 4: In the wild strain 5P14 of *Enterobacter arogenes* ribitol induces gene 1 to synthesize ribitol dehydrogenase. This enzyme metabolizes ribitol, and, with low specific activity, xylitol. Starvation for ribitol leads to a mutation of gene 2 which causes the expression of gene 1 even in the absence of ribitol. Hypermutations of gene 1 and natural selection result in a more specific enzyme.

4.1 hours. Mutations in gene 1 produce strains X2 in 1.7 hours, and then X3 in 0.9 hours. These strains synthesize modified enzymes with increasing specific activity on xylitol.

So a change of the environment (the removal of ribitol in the presence of xylitol) results in specific mutations of two specific genes. The specificity is so great that the experiment is repeatable!

7 Sex

Mate the largest dogs of different litters for several generations, and you end up with huge Great Danes. This is artificial selection. A Great Dane has many "big-dog-genes". These big-dog-genes were already in the genetic pool of the population of dogs. The largest dog of a litter probably has more big-dog-genes than each parent due to the mixing of genes during sexual reproduction (meiosis). The smallest dog of the litter probably has less big-dog-genes than each of the parents.

Natural selection can work in a similar way. An ecological niche attracts individuals specially adapted to that niche. [2] These individuals have an

[2]Mimetism is an example: a green insect that chooses a green environment to hide in

enhanced probability of mating with each other. If the population has a gene pool with several genes that favor the niche, then, after a few generations, these genes can come together and we obtain individuals specially adapted to the niche. If these individuals no longer mate and reproduce with the general population, then a sub-species has formed.

Even cultural preferences can bias mating, resulting in genomes specially adapted to these cultural preferences. This is known as the Baldwin effect, after the description of this phenomenon by James Mark Baldwin in 1896.

8 Viruses

A virus is composed of genetic material packaged (mostly) in proteins. The genetic material may be linear or circular, single or double stranded, haploid or diploid, monopartite or multipartite, DNA and/or RNA. The RNA can be "positive" and serve as a messenger RNA to directly synthesize proteins (using the tRNA and ribosomes of the host cell), or it can be "negative" and require a transcription to +RNA before protein synthesis.

What proteins are coded by the DNA or RNA? In order to reproduce, the virus must code the proteins that form part of the virus itself: structural proteins and enzymes needed prior to protein synthesis (such as enzymes used by retroviruses to synthesize DNA from RNA, enzymes used by negative strand RNA viruses to transcribe -RNA to +RNA, enzymes used by double stranded RNA viruses to make single strands, etc). Depending on the type of virus, other proteins may be coded as well. Examples are enzymes to transcribe +RNA to -RNA, DNA to RNA at various starting sites, proteins that block defenses of the host cell, proteins that cleave other proteins at special sites (so one mRNA of the virus can code many proteins linked together at cleaving sites).

In addition, and of particular interest for evolution, the virus may encode proteins that can turn on or turn off hypermutations, and enzymes used to recombine RNA within the virus, among different viruses (even of different species), and between the virus and the host cell. Let us quote from [6]: "The two major forces acting upon viral genomes to generate diversity that can be tested for environmental survival and replicative fitness are mutations and recombination. Some viruses have a good deal of control over their own rates of mutation and even the frequency of recombination. They exert control by encoding viral enzymes" for replicative and recombinational functions.

has a better chance of survival.

9 Other examples

A strain (known as FC40) of *Escherichia coli* can not digest lactose due to a frameshift mutation in gene *lacZ* that does not allow the synthetis of β-galactosidase in sufficient quantity. It has been observed that this frameshift mutation undergoes reversion when lactose becomes the sole source of energy. It is important to note that most reversions occur after exposure to lactose.[12]

Starvation of *Escherichia coli* induces the production of alternative polymerase enzymes (DinB and $\mathrm{UmuD}'_2\mathrm{C}$) which are capable of replicating badly damaged sequences of DNA, and, in the process, produce high rates of mutations. Homologs of these alternative enzymes have been found in *Saccharomyces cerevisiae*, mice and humans.[13]

Escherichia coli is able to mutate even when not dividing or replicating its DNA, and these mutations may be its main source of genetic variation.[12]

Some plants switch from asexual proliferation (rhizomes) to sexual reproduction in conditions of stress. In doing so they speed up evolution by trying new combinations of genetic material in the process of meiosis.

Snails switch from hermaphrodite reproduction to bi-sexual reproduction in conditions of stress due to parasites. This strategy speeds up evolution when needed.

10 Conclusions

Evolution appears to be hopelessly improbable unless random mutations are limited to no more than about 100 bases of specific genes, and the selection of the outcomes are sufficiently specific. We therefore propose that all living organisms, in addition to being self-organizing and reproducing (autopoyetic), are built to evolve in selective ways. It appears that viruses, prokaryots and eukaryots have considerable control over the rates and spectrum of mutations and recombinations. There are built-in mechanisms to control hypermutations of selective regions of selective genes, and sufficiently selective mechanisms to choose the outcome of these mutations. The high selectivity of these and other mechanisms are required for evolution to be successfull. We have given examples of viruses, prokaryots, eukaryots, and multicellular organisms, where this is indeed the case.

References

[1] The Genome Sequencing consortium, "Initial sequencing and analysis of the human genome", Nature **409**, p. 860 (2001).

[2] Robin D. Knight and Laura F. Landweber, "The Early Evolution of the Genetic Code", Cell, **101**, p. 569 (2000).

[3] Massimo Di Giulio, "The origin of the tRNA molecule: implications for the origin of protein synthesis", J. of Theoretical Biology, **226**, p. 89 (2004).

[4] Scott M. Stagg, Ashley A. Frazer-Abel, Paul J. Hagerman and Stephen C. Harvey, "Structural Studies of the tRNA Domain of tmRNA", J. Mol. Biol., **309**, p. 727 (2001).

[5] Henry M. Sobell, "Molecular Mechanism for Genetic Recombination", Proc. Nat. Acad. Sci. USA, **69**, p. 2483 (1972).

[6] Bernard N. Fields *et.al.*, "Fields Virology", Lippincott-Raven Publishers (1996).

[7] Barbara E. Wright, "A Biochemical Mechanism for Nonrandom Mutations and Evolution", Journal of Bacteriology, **182**, No. 11, p. 2993 (2000), and references therein.

[8] Abul K. Abbas, Andrew H. Lichtman, Jordan S. Pober, "Cellular and Molecular Immunology", fourth edition, W.B. Saunders Co., (2000).

[9] Myron F. Goodman and Matthew D. Scharff, "Identifying protein-protein interactions in somatic hypermutation", JEM **201**, p. 493 (2005).

[10] S. A. Lerner, T. T. Wu, E. C. C. Lin, "Evolution of a Catabolic Pathway in Bacteria", Science, **146**, p. 1313 (1964).

[11] Wu, T. T., E. C. C. Lin, and S. Tanaka, "Mutants of *Aerobacter aerogenes* capable of utilizing xylitol as novel carbon.", J. Bacteriol. **96**, p. 447 (1968).

[12] John Cairns and Patricia L. Foster, "Adaptive Reversion of a Frameshift Mutation in *Escherichia coli*", Genetics **128**, p. 695 (1991).

[13] David Metzgar and Christopher Wills, "Evidence for the Adaptive Evolution of Mutation Rates", Cell, **101**, p581 (2000).